DOWNWARD HIERARCHICAL CLASSIFICATION OF MULTIVALUE DATA

The present invention relates to a method of classifying data in a descending hierarchy, each datum being associated with particular initial values of attributes that are common to the data. More particularly, the invention relates to a method of classification comprising recursive steps of sub-dividing data sets.

BACKGROUND OF THE INVENTION

The Williams & Lambert method of automatic classification is a method of this type. Nevertheless, it applies to data having attributes that are binary, i.e. attributes that for each datum take a particular "true" or "false" value. In that method, on each step of sub-dividing a set, the chi2 value accumulated over all of the other attributes is calculated for each attribute (where the chi2 value calculated between two attributes enables the linkage between those two attributes to be estimated). Thereafter, the set is subdivided into subsets on the basis of the attribute having the greatest accumulated chi2 value.

That method can be extended to classifying data having attributes that take symbolic values, providing a preliminary step of "binarization" is performed. During this step, each symbolic value that an attribute can take is transformed into a binary attribute. Thereafter, during the recursive steps of subdivision, chi2 values are calculated on contingency matrices of the resulting pairs of binary attributes.

However, that method cannot be applied without major drawbacks to classifying multivalue data comprising a mixture of numerical and symbolic

attributes, i.e. data in which some of the attributes are symbolic and other attributes are numerical. In the present document, values are said to be "numerical" when they constitute quantitative values (represented by numbers) and values are said to be "symbolic" when they represent qualitative values (also know as discrete values, e.g. suitable for being represented by letters or words).

For numerical attributes, preliminary discretization of the values over intervals is required so as to make each numerical attribute symbolic. Unfortunately, that transformation inevitably causes information to be lost, without taking into account the fact that the number of discretization intervals will have an influence on the final result, and without it being possible to make a judicious selection of said number of intervals a priori. This affects the coherence of the resulting classes.

In addition, even with attributes that are purely symbolic, the preliminary step of "binarization" considerably increases the number of attributes, thereby considerably increasing the time required to perform the method.

Finally the chi2 calculation is an estimate of the linkage between two attributes, showing up attributes that are correlated or anti-correlated. That calculation thus artificially overestimates the linkage between anti-correlated attributes that result from the binarization step. Since chi2 calculation is also symmetrical between two variables, it does not make it possible to determine whether one variable is more discriminating than another.

SUMMARY OF THE INVENTION

The invention seeks to remedy those drawbacks by providing a method of classification into a descending hierarchy that is capable of treating multivalue data that are numerical and/or symbolic while optimizing the complexity of the treatment and the coherence of the resulting classes.

The invention thus provides a method of classifying data in a descending hierarchy, each datum being associated with particular initial values of attributes that are common to the data, the method comprising recursive steps of subdividing data sets, and wherein, during each step of subdividing a set, discrete values are calculated for the attributes from the particular initial values of the data attributes of said set, and wherein said set is subdivided into subsets as a function of the discrete values.

While executing a classification method of the invention, new discrete values are calculated for attributes associated with the data that are to be classified at each recursive subdivision step of the method. Since this discretization is not performed once and for all during a preliminary step, no information is lost while executing the method. In addition, on each iteration, a set is subdivided into subsets on the basis of the discrete values for the attributes as calculated on a temporary basis, and as a result the method is simplified.

Optionally, during each step of subdividing a set, binary attribute values are calculated from the particular initial attribute values of the data of said set, and said set is subdivided into subsets as a function of the binary values.

This principle of making each numerical and symbolic attribute discrete on only two values ("binarization") maximizes the speed with which the algorithm executes without significantly harming its precision on large volumes of data.

A classification method of the invention may further comprise one or more of the following features:

- during the step of calculating the binary values for the attributes, for each attribute that is numerical, the median value of the particular initial values of said attribute in the data of said set is estimated, and the value "true" is given to the binary attribute corresponding to said attribute for a datum of said set if the particular initial value of the numerical attribute of said datum is less than or equal to the estimated median value, else the value "false" is given thereto;

- the estimated median value of a numerical attribute is obtained as follows:

- extracting extreme values from the set of values taken by the numerical attribute for the data of said set;

- calculating the mean of the remaining values; and

- allocating the value of said mean as the estimated median value;

- during the step of calculating the binary values for the attributes, for each attribute that is symbolic the modal value of the particular initial values of said attribute in the data of said set is estimated, and the value "true" is allocated to the binary attribute corresponding to said attribute for a datum of said set if the initial particular value of the symbolic attribute of said datum is equal to the estimated modal value, else the value "false" is given thereto;

- the modal value of a symbolic attribute is estimated as follows:

- the first m different symbolic values taken by the data of said set for the symbolic attribute are stored, where m is a predetermined number;

- the symbolic value that appears most frequently is retained, amongst said m first different symbolic values; and

- the retained symbolic value is used as the estimate of the modal value;

- said set is subdivided into subsets as a function of a homogeneity criterion calculated on the basis of the discrete values for the attributes of said set;

- said set is subdivided on the basis of the discrete values of the most discriminating attribute, i.e. the attribute for which a homogeneity criterion for all of the discrete values of the other attributes in the resulting subsets is optimized;

- for any attribute, the homogeneity criterion is an estimate of the expectation of the conditional probabilities for correctly predicting the other attributes, given knowledge of this attribute; and

- for certain attributes marked a priori as being "taboo" by means of a particular parameter, the attribute considered as being the most discriminating is the attribute that is not marked as being taboo for which the homogeneity criterion for all of the discrete values of the other attributes in the resulting subsets is optimized.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be better understood from the following description given purely by way of example and made with reference to the accompanying drawings, in which:

- Figure 1 is a diagram showing the structure of a computer system for implementing the method of the invention, and also the structure of the data input to and output by the system; and

- Figure 2 shows the successive steps of the method in accordance with the invention.

DETAILED DESCRIPTION OF THE INVENTION

The system shown in Figure 1 is a conventional computer system comprising a computer 10 associated with random access and read-only type memories RAM and ROM (not shown), for storing data 12 and 14 as input to the computer 10 and as output from the computer 10. The data 12 input to the computer 10 is, for example, stored in the form of a database, or merely in the form of a single file. The data output by the computer 10 is stored in a format making it possible, for implementation of the method of the invention, to represent the data in the form of a tree structure, such as a decision tree 14.

The data 12 is multivalued numerical and/or symbolic data. By way of example, the data may come from a medical or a marketing database, i.e. a database that generally contains several millions of records each associated with several tens of numerical or symbolic attributes.

In the description below, the set of data is written $D = \{d_1, ..., d_n\}$. The set of attributes is written $A = \{a_1, ..., a_p\}$. Thus, each multivalued datum d_i can be represented in attribute space A in the following form:

$d_i = (a_1(d_i); ..., a_p(d_i))$, where $a_j(d_i)$ is the value taken by attribute a_j for datum d_i.

The attributes a_j may be numerical or symbolic. For example, as shown in Figure 1, attribute a_1 is numerical. It takes the value 12 for datum d_1 and the value 95 for datum d_n. Attribute a_p is symbolic. By way of example, it allocates a color to the database: thus, datum d_1 is of color blue and datum d_n is of color red.

It is judicious to represent this multivalued database in the form of a table in which each row corresponds to one datum d_i and in which each column corresponds to one attribute a_j.

The computer 10 implements an automatic classification method for classifying the multivalued numerical and/or symbolic data 12 into a descending hierarchy, for the purpose of generating homogeneous classes of data, which classes are accessed with the help of the associated decision tree 14.

A preferred implementation of the invention is to organize the resulting classes into a binary decision tree, i.e. an implementation in which any one data class is subdivided into two subclasses. This particularly simple implementation enables data to be classified quickly and efficiently.

To implement the classification method, the computer 10 has a driver module 16 whose function is to coordinate activation of an input/output (I/O) module 18, a

discretization module 20, and a segmentation module 22. By synchronizing these three modules, it enables the decision tree 14 and homogeneous classes to be generated recursively.

The function of the I/O module 18 is to read the data 12 input to the computer 10. In particular, its function is to identify the number of data to be processed and the types of the attributes associated with data, in order to supply them to the discretization module 20.

The function of the discretization module 20 is to transform the attributes a_1, ..., a_p into discrete attributes. More precisely, in this example, the discretization module 20 is a binarization module having the function of transforming each attribute into a binary attribute, i.e. an attribute that can take on only the value "true" or the value "false" for each of the data d_i. Its operation is described in detail below with reference to Figure 2.

The function of the segmentation module 22 is to determine from the binary attributes calculated by the binarization module 20, which attribute is the most discriminating for subdividing a data set into two subsets that are as homogeneous as possible. Its operation is described in detail below with reference to Figure 2.

The recursive method of automatic classification and of generating an associated decision tree comprises a first step 30 of extracting data from the database 12. During this step, data belonging to a set E_1 are extracted from the database 12, said set being represented by a terminal node of the decision tree 14, and being for subdivision into two subsets E_{11} and E_{12}.

The data are extracted together with their attributes and the latter are delivered to the input of the binarization module 20 which processes symbolic attributes and numerical attributes separately.

Thus, during a step 32a of estimating a median value, the binarization module 20 calculates for each numerical attribute a_j, an estimate of the median value of the following set of values:

$\{d_1(a_j); ...; d_n(a_j)\}$

During this step 32a, it is possible to calculate the median value m_j of the set of values taken by the attribute a_j directly, however such a calculation can be replaced by a method of estimating this median value, which method is easier to implement by computer means.

This method of estimating the median value M_j comprises the following steps, for example:

- the extreme values of the set of values taken by the attribute a_j are extracted;
- the mean of the remaining values is calculated; and
- M_j is given the value of this mean.

The extreme values extracted from the set are constituted, for example, by the $\underline{n}$ largest values and the $\underline{n}$ smallest values, where $\underline{n}$ is a predetermined parameter or is the result of earlier analysis of the distribution of the values taken by the attribute a_j.

It is also possible to estimate the median value merely by calculating the mean of all of the values of the attribute.

During the following step 34a of calculating binary attributes, values are calculated for a binary attribute b_j on the basis of each numerical attribute a_j as follows:

if $d_i(a_j) \leq M_j$, then $d_i(b_j)$ = true

if $d_i(a_j) > M_j$, then $d_i(b_j)$ = false

For the symbolic attributes a_k, the binarization module 20 calculates for each of them an estimate of the modal value of their values. This is implemented during a modal value estimation step 32b.

The modal value M_k of a set of symbolic values for an attribute a_k is the symbolic value that this attribute takes most often.

The modal value M_k can be calculated, however that is expensive in terms of computation time.

In order to simplify this step, direct calculation of the modal value can be replaced by a method of estimating it, which method comprises the following steps:

- while reading the data of the set E_1, the binarization module 20 stores the $\underline{m}$ first different symbolic values taken by the data d_i for the attribute a_k, where $\underline{m}$ is a predetermined number;

- the symbolic value which appears most often is retained, amongst said $\underline{m}$ first different symbolic values; and

- this retained symbolic value is allocated to the modal value M_k.

By way of example, $\underline{m}$ is selected to be equal to 200.

If the number of possible symbolic values for the attribute a_k is less than $\underline{m}$, then the estimated modal value M_k is equal to the modal value itself. Otherwise, the estimated modal value M_k is highly likely to constitute a good replacement value for the modal value in many cases. In general, most symbolic statistical attributes have fewer than several tens of different symbolic values.

During following step 34b for calculating binary attributes, the values of a binary attribute b_k are calculated from each symbolic attribute a_k as follows:

if $d_i(a_k) = M_k$, then $d_i(b_k) = \text{true}$

if $d_i(a_k) \neq M_k$, then $d_i(b_k) = \text{false}$

Following steps 34a and 34b, the method moves on to a step 36 during which the binary attributes b_k, b_j derived from the symbolic attributes a_k and the numeric attributes a_j are reassembled. This constitutes a set $B = \{b_1, ..., b_p\}$ of binary attributes for the set E_1 of data d_i. During this step, the binarization module 20 supplies the multivalue data of the set E_1 associated with their binary attributes $\{b_1, ..., b_p\}$ to the segmentation module 22.

Thereafter, during a calculation step 38, the segmentation module 22 calculates for each attribute b_j the following value $f(b_j)$:

$$f(b_j) = \sum_{k,k\neq j} FU(b_j,b_k)$$

where:

$$FU(b_j,b_k) = \frac{1}{n}[c(B_j)\text{Max}(p(B_k/B_j);p(\neg B_k/B_j)) + c(\neg B_j)\text{Max}(p(B_k/\neg B_j);p(\neg B_k/\neg B_j)]$$

where:

for all index j, B_j is the event "the attribute b_j takes the value true"; and $\neg B_j$ is the event "the attribute B_j takes the value false";

with Max(x,y): the function that returns the maximum of $\underline{x}$ and $\underline{y}$;

p(x/y): the probability of event $\underline{x}$, given knowledge of the event $\underline{y}$; and

c(x): the number of instances of event $\underline{x}$ (weighting).

As described above, for each attribute b_j, the value $f(b_j)$ is an estimate of the expectation that conditional probabilities will correctly predict the other attributes, knowing the value of the attribute b_j. In other words, it makes it possible to evaluate the pertinence of segmentation into two subsets based on the attribute b_j.

Nevertheless, some other function $\underline{f}$ could be selected for optimizing segmentation, such as a function based on calculating the covariance of attributes.

During following selection step 40, the segmentation module 22 determines the binary attribute b_{jmax} which maximizes the value of $f(b_{jmax})$, i.e. the attribute which is the most discriminating for segmentation into two subsets.

Thereafter, during a segmentation step 42, the module 22 generates two subsets E_{11} and E_{12} from the data set E_1. The first subset E_{11} is constituted, for example, by a subset combining all of the data for which the attribute b_{jmax} takes the value true and the second subset E_{12} groups together all the data of the set E_1 for which the attribute b_{jmax} takes the value false.

During this step, the decision tree 14 is updated by adding two nodes E_{11} and E_{12} connected to the node E_1 by two new branches.

Thus, when moving through this decision tree and on reaching the node E_1, the following test is performed:

"for datum d_i, is the attribute a_{jmax} of a value less than M_{jmax}?", if a_{jmax} is a numerical attribute; or

"for datum d_i, is the attribute a_{jmax} of a value equal to M_{jmax}?", if a_{jmax} is a symbolic attribute.

If the response to this test is positive, then datum d_i belongs to subset E_{11}, else it belongs to subset E_{12}.

Following step 42, during a test step 44, a criterion for stopping the method is tested. This stop criterion is constituted, for example, by the number of terminal nodes in the decision tree, i.e. the number of classes that have been obtained by the classification method, assuming some fixed number of classes not to be exceeded has been previously established.

The stop criterion could also be the number of levels in the decision tree. Other stop criteria could equally well be devised.

If the stop criterion is reached, then the method moves on to an end-of-method step 46. Otherwise it loops back to step 30 and restarts the above-described method on a new data set, for example the set E_{11} or the set E_{12} as previously obtained.

It should be observed that the above-described classification method is a method that is not supervised.

The classification method can also be used in a "semi-supervised" mode. It is useful to apply the classification method in a semi-supervised mode when it is desired

to predict or explain a particular attribute as a function of all the others while this particular attribute is badly or sparsely entered in the database 12, i.e. when a large number of data d_i have no value corresponding to this attribute. Under such circumstances, it suffices to identify this attribute as being purely "to be explained", and to mark it as such via special marking, for example in an associated parameter file. This attribute which is specified as being "to be explained" by the user is referred to as a "taboo" attribute. The taboo attribute must not be selected as a discriminating attribute.

It should also be observed that a plurality of taboo attributes can be defined. Under such circumstances, it suffices to distinguish among the attributes a_j those attributes which are said to be "explanatory" and those attributes which are said to be "taboo". Taboo attributes are then not selected as discriminating attributes when performing segmentation during above-described step 40.

In semi-supervised mode, during step 40, if the selected attribute is a taboo attribute, then a search is made for the second attribute which maximizes the function $f(b_j)$, and so on until the most highly discriminating non-taboo attribute has been found, i.e. the attribute which maximizes the uniformity criterion for the discretized values of the other attributes in the subsets E_{11} and E_{12}.

The classification as finally obtained can subsequently be used for predicting the values of a taboo attribute for data where the values are missing. The classification method performs tests only on those attributes that are explanatory, while taking maximum advantage of all of the correlations between attributes.

Values for a taboo attribute are predicted by replacing the values that are missing or sparsely entered by the most probable values that are given in each class.

It can clearly be seen that a method of the invention enables classification to be performed simply and efficiently in a descending hierarchy on multivalued numerical and/or symbolic data. Its low level of complexity makes it a suitable candidate for classifying large databases.